Hierarchical interpretations for neural network predictions

21 Jan 2020

Presenter: Sanchit Sinha

https://qdata.github.io/deep2Read/

Motivation

- Explanation of DNNs outputs and predictions required for several purposes included reducing bias, regulatory, etc.
- **Bottom up - Agglomerative Clustering** type approach which is the most intuitive to general users can be used to explain DNN predictions
- Can be used to **compare models** with similar output - the ones with better explanation can be used with confidence
- Explanation methods should have **robustness** against adversarial perturbations
- Can give **hierarchical saliency** - identify parts which have the maximum influence in predictions

Background

- Contextual Decomposition: Assigns scores to **phrases** in an LSTM setting to measure their importance with respect to the actual prediction made

$$\beta_t^c = \beta_t^f + \beta_t^u$$
$$\gamma_t^c = \gamma_t^f + \gamma_t^u$$

- CD is limited to memory networks, need to expand it to CNNs.
- **Locality importance has limitations** - no idea what the intermediate layers of the DNN is learning. For eg. word level score although important is not the ultimate way of deciding polarity o sentiment
- **Agglomerative Hierarchy** is a good intuitive way of understanding explanations

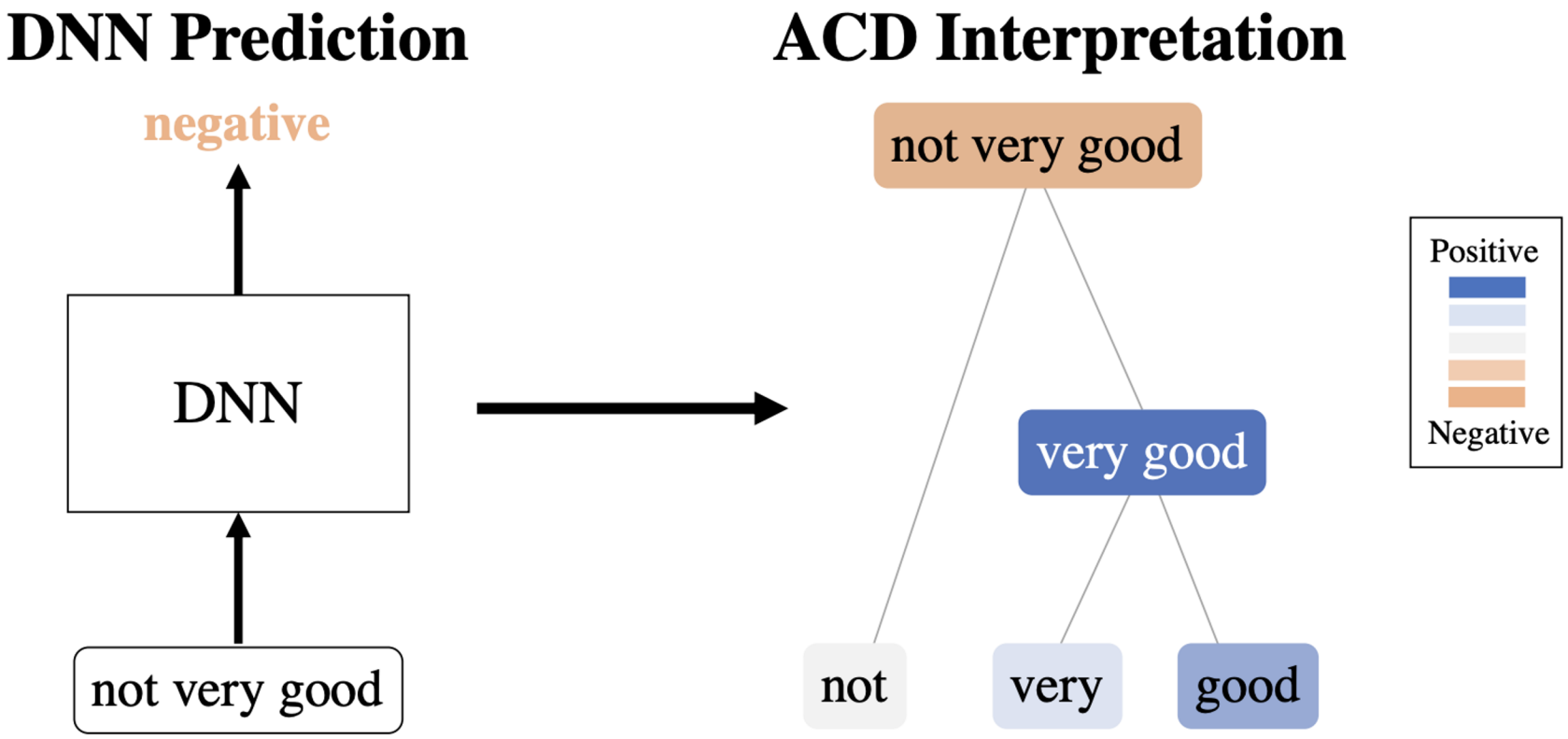
Related Work

- W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. arXiv preprint arXiv:1801.05453, 2018.
- Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klaus-Robert Muller, ¨ and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.
- W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592, 2019.

Claim / Target Task

1. Generalize the CD score calculation to CNNs with convolutional layers and max-pooling layers
2. Use the CD scores as measures to build explanations in a bottom up way kind of like clustering
3. The CD scores are used to combine the candidates and generate new candidates at every level
4. Make algorithm explain the phrase level sentiment
5. Make algorithm explain bias in datasets and ensure robustness against adversarial perturbations

An Intuitive Figure Showing WHY Claim



Proposed Solution

- Expanding CD to general multilayer DNN:

$$g^{CD}(x) = g_L^{CD}(g_{L-1}^{CD}(...(g_2^{CD}(g_1^{CD}(x)))))$$

- Expressing contribution of every layer importance as sum of feature set and non-feature set.
  - For a convolution layer:

$$\beta_i = W\beta_{i-1} + \frac{|W\beta_{i-1}|}{|W\beta_{i-1}| + |W\gamma_{i-1}|} \cdot b$$
$$\gamma_i = W\gamma_{i-1} + \frac{|W\gamma_{i-1}|}{|W\beta_{i-1}| + |W\gamma_{i-1}|} \cdot b$$

  - For max-pool layers:

$$max\_idxs = \underset{idxs}{\text{argmax}} \ [\text{maxpool}(\beta_{i-1} + \gamma_{i-1}; idxs)]$$
$$\beta_i = \beta_{i-1}[max\_idxs]$$
$$\gamma_i = \gamma_{i-1}[max\_idxs]$$

  - For ReLU:

$$\beta_i = \text{ReLU}(\beta_{i-1})$$
$$\gamma_i = \text{ReLU}(\beta_{i-1} + \gamma_{i-1}) - \text{ReLU}(\beta_{i-1})$$

Implementation

Algorithm 1 Agglomeration algorithm.

```
ACD(Example x, model, hyperparameter k, function CD(x, blob; model))
    # initialize
    tree = Tree()                                                    # tree to output
    scoresQueue = PriorityQueue()                                    # scores, sorted by importance
    for feature in x :
        scoresQueue.push(feature, priority=CD(x, feature; model))

    # iteratively build up tree
    while scoresQueue is not empty :
        selectedGroups = scoresQueue.popTopKPercentile(k)            # pop off top k elements
        tree.add(selectedGroups)                                     # Add top k elements to the tree

        # generate new groups of features based on current groups and add them to the queue
        for selectedGroup in selectedGroups :
            candidateGroups = getCandidateGroups(selectedGroup)
            for candidateGroup  in candidateGroups :
                scoresQueue.add(candidateGroup, priority=CD(x, candidateGroup;model)-CD(x,selectedGroup;
    model))
    return tree
```

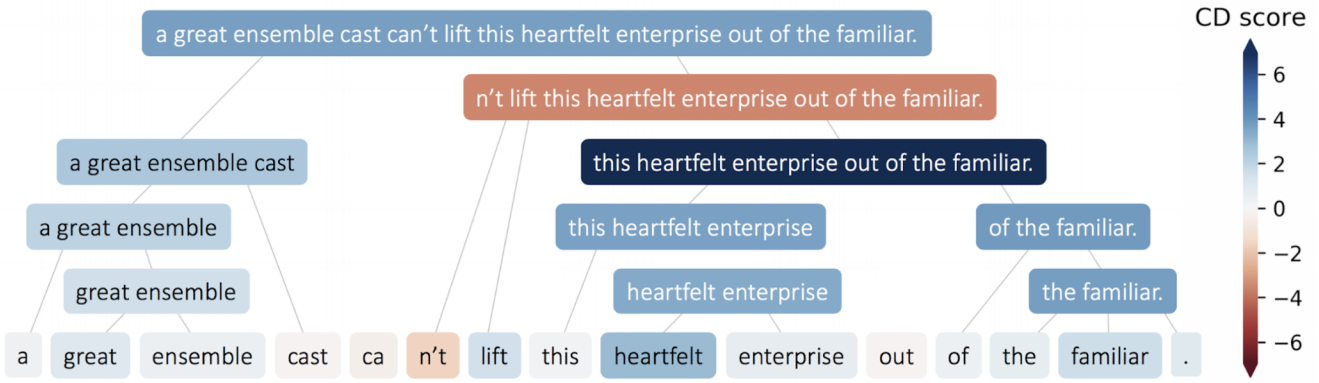
Data Summary

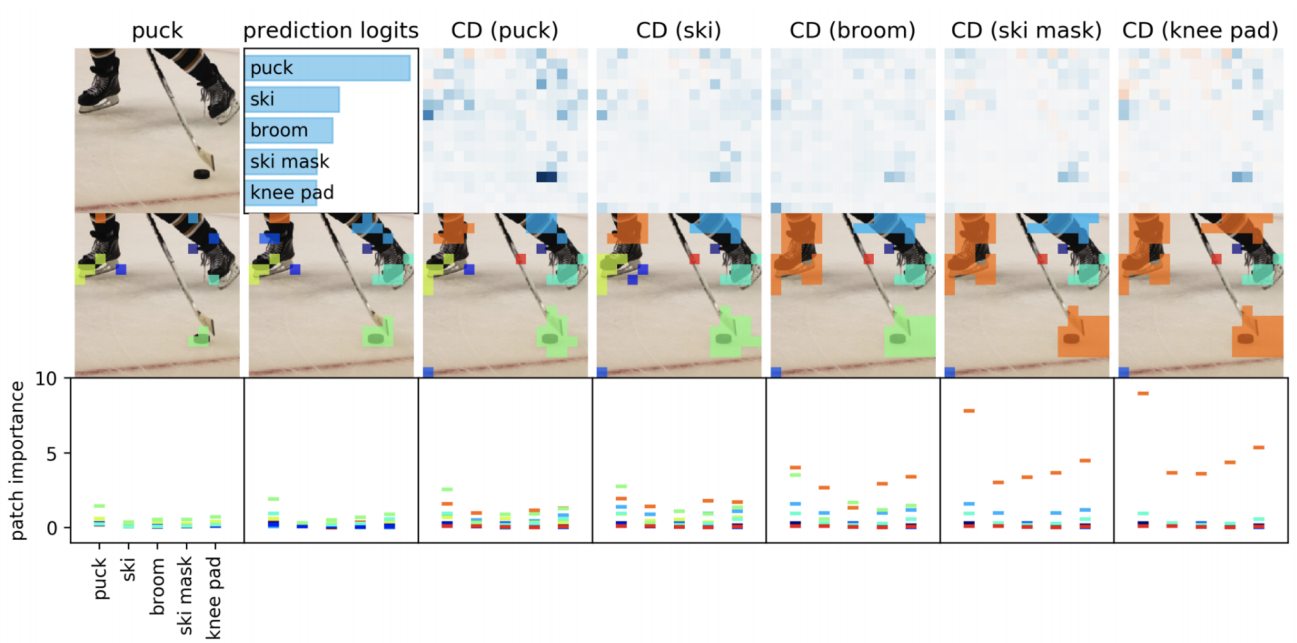
Imagenet

MNIST

SST

Experimental Results

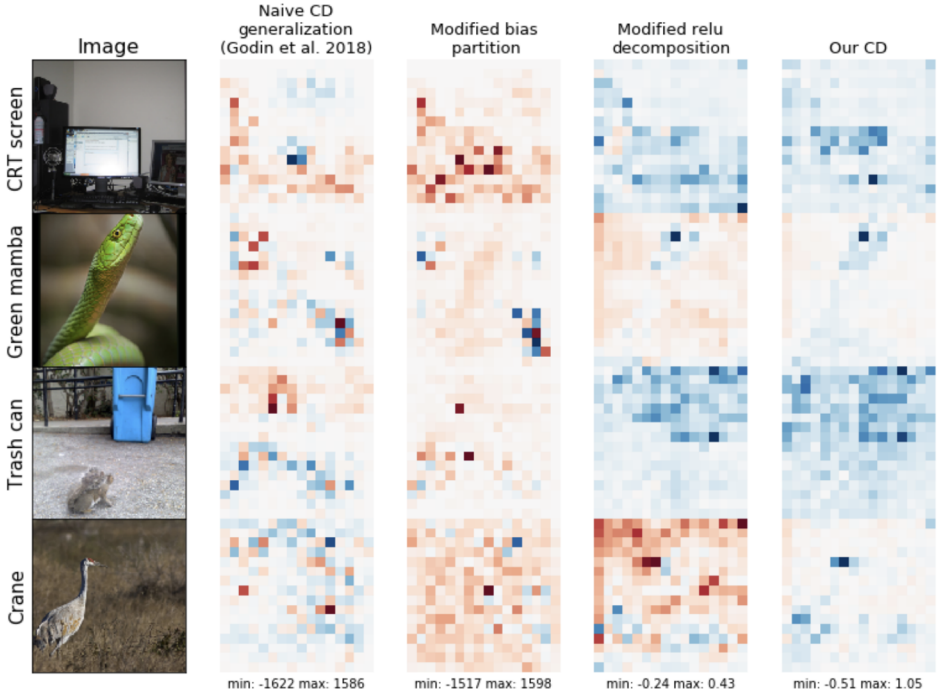


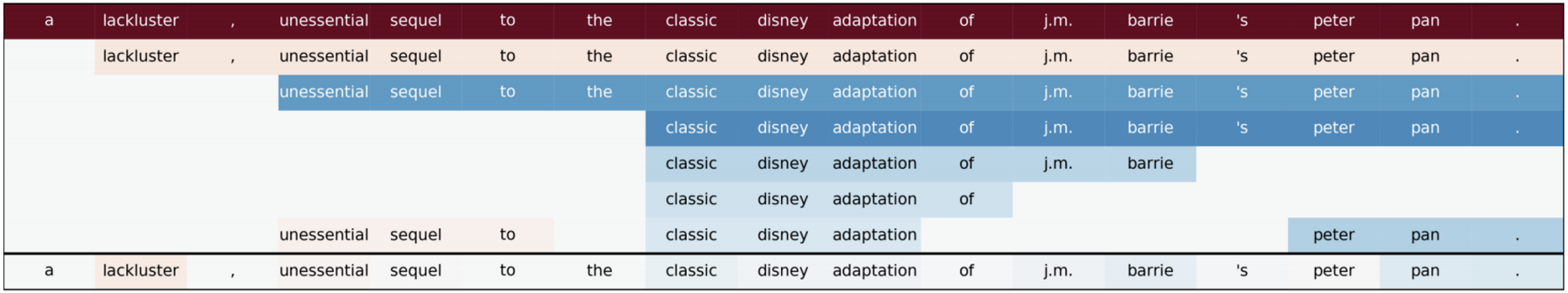


Attack Type	ACD	Agglomerative Occlusion
Saliency (Papernot et al., 2016)	0.762	0.259
Gradient attack	0.662	0.196
FGSM (Goodfellow et al., 2014)	0.590	0.131
Boundary (Brendel et al., 2017)	0.684	0.155
DeepFool (Moosavi Dezfooli et al., 2016)	0.694	0.202

Experimental Analysis

- The hockey and puck example highlights a glaring bias in the system. The detection of puck is dependent upon the presence of skates in the system.
- Now if a similar object is replaced in position of puck with same backgrounds and settings, the model will tend to call it a puck




a lackluster , unessential sequel to the classic disney adaptation of j.m. barrie 's peter pan .
lackluster , unessential sequel to the classic disney adaptation of j.m. barrie 's peter pan .
unessential sequel to the classic disney adaptation of j.m. barrie 's peter pan .
classic disney adaptation of j.m. barrie 's peter pan .
classic disney adaptation of j.m. barrie
classic disney adaptation of
unessential sequel to classic disney adaptation peter pan .
a lackluster , unessential sequel to the classic disney adaptation of j.m. barrie 's peter pan .

Conclusion/ Future Work

- Can identify bias in datasets
- Clear explainability as a bottom up prediction
- Robustness to adversarial perturbations
- Paving way for intuitive explainability

Plan for Sanchit

- Taking a step back and reading about saliency map generations - Guided Backprop and GradCAM
- Reading if Saliency methods can be applied to GraphNN

References

- W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. arXiv preprint arXiv:1801.05453, 2018.
- Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klaus-Robert Muller, ¨ and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7):e0130140, 2015.
- W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592, 2019.